A LANGUAGE MODELING APPROACH TO DETECT BIAS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Ceren ATİK

July, 2020
İZMİR

ACKNOWLEDGMENTS

I would like to thank my advisor Asst. Prof. Dr. Selma Tekir, who always shed light on my way with her knowledge, ideas and guidance. I am grateful that I have the opportunity to work with her. In this process, she was always patient, understanding and positive.

I would like to thank Asst. Prof. Dr. Nesli Erdoğmuş and Asst. Prof. Dr. Zerrin Işık for accepting to be a part of my thesis defence jury and sparing their valuable time to read and evaluate my thesis and provide me with feedbacks.

Also I would like to thank my family. They always supported me at to do better all times and under all circumstances.

ABSTRACT

A LANGUAGE MODELING APPROACH TO DETECT BIAS

Technology is developing day by day and is involved in every area of our lives. Technological innovations such as artificial intelligence can strengthen social biases that already exist in society, regardless of the developers' intentions. Therefore, researchers should be aware of this ethical issue.

In this thesis, the effect of gender bias, which is one of the social biases, on occupation classification is investigated. For this, a new dataset was created by collecting obituaries from the New York Times website and they were handled in two different versions, with and without gender indicators. Since occupation and gender are independent variables, gender indicators should not have an impact on the occupation prediction of models. In this context, in order to investigate gender bias on occupation estimation, a model in which occupation and gender are learned together is evaluated as well as models that make only occupation classification are evaluated. The results obtained from models state that gender bias has a role in classification occupation.

ÖZET

TARAFLILIĞIN TESPİTİ İÇİN BİR DİL MODELİ YAKLAŞIMI

Teknoloji günden güne gelişerek hayatımızın her alanına dahil olmaktadır. Yapay zekâ gibi teknolojik yenilikler, geliştiricilerin niyetlerinden bağımsız olarak toplumda zaten var olan sosyal önyargıları güçlendirebilir. Bu nedenle, araştırmacılar bu etik sorunun farkında olmalıdır.

Bu tez çalışmasında, sosyal önyargılardan biri olan cinsiyet yanlılığının meslek sınıflandırması üzerindeki etkisi araştırılmaktadır. Bunun için New York Times web sitesinden anma yazıları toplanarak yeni bir veri kümesi oluşturulmuş ve bu anma yazıları cinsiyet göstergeleri dahil ve hariç olmak üzere iki farklı versiyonuyla ele alınmıştır. Meslek ve cinsiyet birbirinden bağımsız değişkenler olduğu için cinsiyet göstergelerinin modellerin meslek tahmini üzerinde bir etkisi olmadığı varsayılmaktadır. Bu bağlamda, meslek tahmini üzerinde cinsiyet yanlılığını araştırmak için sadece meslek sınıflandırması yapan modellerin yanında meslek ve cinsiyetin aynı anda öğrenildiği bir model de değerlendirilmiştir. Deneysel sonuçlar meslek tahmininde cinsiyet yanlılığının etkili olduğunu ortaya koymaktadır.

LIST OF CONTENTS

APPENDIX

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF ABBREVIATIONS

NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NYT	New York Times
SOC	Standart Occupation Classification

CHAPTER 1

INTRODUCTION

Technological innovations such as artificial intelligence can reinforce social biases that already exist in society, regardless of the developers' intentions. If we examine the term bias, it is generally the disproportionate weight that is created in favor of or against an idea or thing unfairly. Bias can be classified using two basic mental structures: Explicit bias and implicit bias. Explicit bias comes from conscious bias, while implicit bias comes from the subconscious. Implicit bias refers to attitudes or stereotypes that unconsciously affect our understanding, actions and decisions. Bias is difficult to detect and evaluate because it is usually implicit. People can develop bias toward or against an individual, an ethnic group, gender identity. Therefore, researchers should be aware of this ethical issue.

Science and technology are concerned with imitating people's behavior. These behaviors include our conscious or subconscious attitudes. Artificial intelligence refers to the intelligence shown by machines or computer software. As Alan Turing describes in his study of machinery intelligence (Turing 1950), it is a topic that has been debated for many years that machines exhibit human behavior and learn like humans. Deep learning enables the machine to process and make sense of data through an artificial neural network very similar to the human brain. Suppose that all the functions that our brain performs biologically are a kind of deep neural network that gives labels as outputs. If we look at our vision system, when we look at something, we don't just perceive objects. We can handle them in pieces and understand the whole. Just by looking, we can infer the mood of people on the street, how they dress, the weather according to their clothes, which country and even which city it is. On the other hand, modern neural networks and machine learning algorithms usually solve a single problem. A single result is generally expected at the end of operations such as classification. In multitask learning, more than one output is obtained. It is also important because it is close to “human-like” behavior.

In the thesis study, the estimation of gender and occupational variables will be performed from the data collected from the NYT Obituaries website[1]. The articles in the dataset are handled in two ways and they are separated as states containing gender indicators and cleaned from these indicators. The aim is to reveal gender bias on occupational classification. If gender and profession are considered as two independent variables, the results obtained from the articles containing the gender indicators and the articles removed from these indicators should be the same. When an article with gender indicators is given as input to the model, the estimated professional information for women and men should be the same with the model is given articles without gender indicators. If we observe a difference between the two results, we can say that gender information plays a role in the estimation, there is a dependency between the two variables, and therefore gender bias occurs. These classifications will first be carried out using Support Vector Machine (SVM) (Cortes and Vapnik 1995) , Hierarchical Attention Networks (HAN) (Yang, et al. 2016), and DistilBERT (Sanh, et al. 2019) which are accepted in the literature. Afterwards, multi-task learning approach and classification tasks will be tried to be learned jointly. The purpose of addressing the multitask learning model is to define a separate variation at the point of evaluating the interactions of the variables when the variables are learned together. In this way, the interaction of variables will be systematically evaluated. The study aims to detect gender bias by using machine learning and deep learning models as single task and multi-task learning. The progress of technology day by day and the inclusion of it to many areas of our lives have brought a new dimension to these social problems. In the related work section, the effect of this dimension on different fields and studies is examined. Experimental design and methodology section introduces our models and methodology to detect gender bias and also gives an information about our dataset and how to preprocess it. Some background information about models used is given in background section. In experimental result section, the results obtained from the models used and whether there are signs of gender bias in the results are discussed. In the last section, we conclude our results and some possible issues on future work and research on this subject.

[1] https://www.nytimes.com/section/obituaries

CHAPTER 2

BACKGROUND

In this chapter, some fundamental concepts of the models used in the thesis will be discussed. The TF-IDF concept and dimensionality reduction for SVM model and the encoder and attention mechanism for neural networks will be mentioned.

2.1. TF-IDF

TF-IDF (Jurafsky and Martin 2008) is a abbrevation of "term frequency-inverse document frequency". It is a weight calculated by statistical methodology to evaluate the importance of a term in a corpus. TF-IDF consist of two components; term frequency and inverse term frequency. Term frequency is calculated by dividing the selected term by the total number of terms in the text as in equation 2.1.

$$tf(d,t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document\ d} \qquad (2.1)$$

Inverse term frequency computed as the logarithm of the total number of text in corpus divided by the number of text containing the term as in equation 2.2.

$$idf(t) = \ln\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ the\ term\ t} \right) \qquad (2.2)$$

The TF-IDF value is obtained by multiplying $tf(d,t)$ and $idf(t)$ value where t is a term and *d* is a document.

2.2. Dimensionality Reduction

The number of features determines the dimension of the space where the data will be represented. N-features can be expressed in N-dimensional space. Working in high-dimensional space has some difficulties and side effects. As the data size grows, density and distance information become meaningless and this affects the performance of the algorithms. Since these difficulties occur as the size of space increases, the effect created by the size increase is called the curse of dimensionality. Dimensionality reduction techniques such as Singular Value Decomposition (SVD) deal with the "curse of dimensionality" by extracting new features from the data. It allows us to decompose an any large matrix M of relationships into three simpler ones -U, Σ, V- that can be used to approximately reconstruct the original data (Singular Value Decomposition 2020).

$$M = U\Sigma V^T \qquad (2.3)$$

Where M is an mxn matrix, U is an mxm orthogonal matrix, Σ is an mxn nonnegative rectangular diagonal matrix, V is an nxn orthogonal matrix and V^T is transpose of The diagonal values in Σ matrix are known as the singular values of the original matrix M. The columns of the U matrix are called the left-singular vectors of M, and the columns of V are called the right-singular vectors of M.

SVD can be used to calculate a projection of a dataset and select a number of dimensions. Equation 2.3 can be also written as;

$$M = \sum_{i=1}^{p} \sigma_i u_i v_i^*$$

Where σ_i are the singular values and p is the number of singular values that are non-zero. The u_i ve v_i are ith columns of U and V respectively.

We can form an approximation to M by truncating the sum above. Instead of taking all the singular values, and their corresponding left and right singular vectors, we only take the k largest singular values and their corresponding vectors. This is called the truncated SVD.

2.3. Encoder and Attention Networks

The first neural network model used in this thesis is Hiearchical Attention Network (HAN). This model includes encoder and attention mechanism. Encoder takes in an input sequence and output a context vector. In our case the encoder is a bidirectional GRU (Cho, et al. 2014) and context vector is the concatenation of both directions' final hidden states. GRU has a gated unit that decides whether the flow of information will pass through the unit or not by doing so it passes relevant information down the long chain of sequences to make predictions. After encoding layer, there is an attention layer. An attention mechanism calculates the dynamic weights representing the relative importance of the inputs in the sequence (the keys) for output (the query). Then multiplying these weights with the input sequence (the values), weighted sequence is obtained. Using the sum of weighted vectors, a single context vector is calculated. There are various methods to obtain attention and one of them is Dot-Product Attention. It is basicly dot product of the query and keys. Dot product produces very large vector dimensions which will result in very small gradients when passed into the softmax function so that values prior can be to scaled. Softmax is performed as a normalisation technique to non-linearly scale the weight values between 0 and 1.

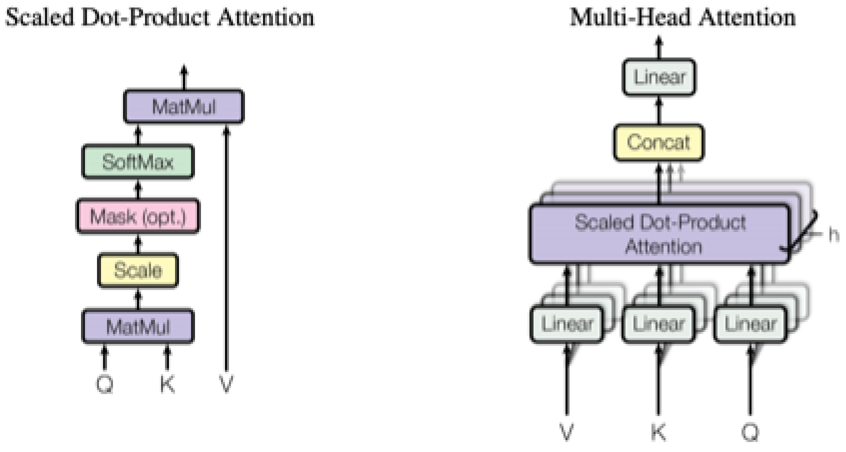


Figure 2.1. Scale Dot Product Attention and Multihead Attention
(Source: *Vaswani, et al. 2017*)

In Figure, multi-head attention is used in transformer models such as BERT (Devlin, et al. 2019). It calculates a representation of the sequence by aligning the words in a sequence with other words in the sequence. The multi-head attention is essentially multiple attention layers jointly learning different representations from different positions. Each calculated context vector is concatenated in this architecture.

CHAPTER 3

RELATED WORK

Detecting bias is becoming gradually important based on its relevance in many fields, ranging from evaluating publications, to sentiment analysis, to business life. In recent years, studies on gender bias in different fields have been examined and these studies have been summarized in the continuation of the text.

If the term bias is associated with people, there are studies that claim algorithms are biased because they are written by people. When lexical sensitivity analysis was applied on the studies written by men and women, a higher accuracy was obtained in the studies of women writers (Thelwall 2018). This study aims to reveal gender bias in sentiment analysis and uses two best-known classifiers for this; Naive Bayes and SVM.

The study tries to find out whether the accuracy value of the models changes according to gender. It tries different cases. One of them is training with one-gender reviews and testing with reviews of same gender as in training. And the other is training with different size corpora includes both genders or single gender and testing with reviews includes both gender or single gender. Scenarios where the training data include both gender and only one gender are tested. The results obtained are compared and it is revealed that the reviews written by women have higher accuracy than the ones written by men. Also it is shown that the size of corpora effects the result. Larger corpora gives better results for all conditions.

It is important to note that bias is not limited to gender and may apply to both genders. People can develop bias against an individual, an ethnic group as well as gender identity. In some studies, a few of them can be seen together. As in the previous study, Buolamwini and Gebru (Buolamwini and Gebru 2018) claim that machine learning algorithms may be gender biased and provide an approach to assess the prejudice present in automated face analysis algorithms and datasets according to phenotypic subgroups. This study also evaluates skin color. Although the results are intended to be used in market preparation, the gender and phenotypic distribution of the results shows that the performance varies for different subgroups. The results show that classifiers generally work better with more light-skinned individuals and men. The results of the studies are important in terms of identifying, addressing the current situation in advertising or other fields and formulating strategies.

Gender bias is preferring one gender over another or prejudicing one gender. We understand that the definition of gender bias refers to prejudice that occurs in an environment where the same conditions are presented. When considering natural language processing to analyze texts, word embedding should be mentioned since there is a risk of increasing bias in data (Bölükbaşı, et al. 2016). The analogies produced from these embeddings reveal bias regarding the data on which they are trained. To measure bias, the authors compared a word vector with vectors of a gender-specific word pair, such as “nurse” and “woman”. The vector of the word “nurse” is also close to the vector of the word “man”, but the distance is unequal and thus causes bias. The aim of this study is to preserve the beneficial features of embedding while reducing gender bias in word embeddings. This study demonstrates the importance of word embeddings in terms of bias used as input in many natural language processing applications.

Wikipedia, one of the most popular online sources of information, is an encyclopedia created by a community, and therefore the Wikipedia editor community has the potential to introduce systemic prejudices such as gender bias into Wikipedia's content (Wagner, et al. 2015). This may lead to an increase in inequalities already present in the real world. The study examines articles about notable people in six different languages in terms of gender bias. As a result, they found that on Wikipedia, men and women were equally well received in all six languages. However, on Wikipedia, they found that the way women are portrayed is different from the way men are portrayed. In this case, there is no evidence of gender bias on Wikipedia. The study provides evidence of gender inequalities.

One of the other gender bias studies (Caplar, Tacchella and Birrer 2017) touches on a very interesting point. The study underlines the role of the first author gender on the number of citations. Male and female authors should receive the same number of citations for articles with gender-independent attributes, such as seniority and number of references, but this study shows the opposite. Male writers receive more citations than female writers. The random forest algorithm is trained through articles in which the first author is a male, then the algorithm is trained to estimate the number of citations expected from the articles in which the first author is female. By doing so, the results are compared and it is revealed that female authors receive fewer citations. Some statistical analyses of the effect of gender on citations were inspired by this study (Thelwall 2018).

As mentioned before, gender bias is encountered not only in academic studies but also in daily life and business life. Another study (Fu, Danescu-Niculescu-Mizil and Lee 2016) reveals gender bias in sports journalism. Interviews with both male and female tennis players show that the questions differ according to gender. While the questions asked to the male players were related to the game, it was seen that the female players were asked the questions which were not relevant. Winning the game and the number of matches played do not change the result. While all the conditions are the same, asking different questions according to gender clearly shows gender bias.

The last study to be examined is distinguished from the previus ones in that it is close to our thesis work. The study (De-Artega, et al. 2019) is about occupational classification which is a multiple classification example. The aim of the study is to reveal gender bias in occupational classification. For this, biographies were collected from the Common Crawl[2] website. Three different models were evaluated after obtaining the gender and occupational knowledge of the individuals through the collected biographies. These are SVM, logistic regression, and deep recurrent neural networks. For each models, two different scenarios are evaluated where gender indicators (name, gendered pronoun) are used and cleared. The aim is to show whether there is a dependency between gender and occupation. The true positive rate obtained from the classification of indicators including gender is compared with the true positive rate obtained from the classification when it is cleared. It is argued that there is gender bias since the value when gender indicators are not used is different from the value when it is used. The study examines only the occupational classification, it does not make gender classification. We use this scrubbing method with different models on a new dataset and also we investigate gender bias in a scenario where occupational and gender variables are learned together.

[2] https://commoncrawl.org

CHAPTER 4

EXPERIMENTAL DESIGN AND METHODOLOGY

As discussed in the related work section, different methods have been used to reveal gender bias. One of them is changing the gender of person and comparing the results and the other is scrubbing gender indicators and comparing results for with and without gender indicators. These two, focuses on genders or gender indicators. There is also another method which is comparing a word vector with vectors of a gender-specific word pair.

In this thesis, it is investigated whether there is gender bias on the data collected from the NYT Obituaries section using scrubbing method with different models. The articles collected to test the hypothesis were preprocessed. Before performing preprocessing, the first sentences of articles were removed. The reason for this is that first sentences contain the name, surname and occupational information of the person and the name usually contains gender information inherently. Not to include the name and the profession in the prediction is a step in the scrubbing method. The same steps will be followed in this thesis. The effect of the first sentence on the prediction is examined in detail in the experimental results section. After the removal of the first sentences, pre-processing was done and the articles were prepared to be input to the models. The articles were tested with different models for gender bias analysis. Our starting point is that gender and profession are independent variables. In this case, the occupation prediction made by models through articles and scrubbed articles for individuals should be the same. If we observe a difference between the two results, we can say that gender information plays a role in the estimation, there is a dependency between the two variables, and therefore gender bias occurs.

4.1. Data Collection

Dataset is one of the basic building blocks of the thesis. In the literature, studies examining gender bias have been found to collect their own datasets. The main reason for this is that there is no dataset in the field to be examined. In this context, the dataset[3] to be used in the thesis study was collected from the NYT web site by web scraping from obituaries section. Although the NYT API was available, this method was used because the entire article and the gender of the person mentioned in the article were not provided. Articles published between 2014 and September 2019 were collected.

The dataset collected with NYT API includes article, article summary, title, author, the date of publication, category, the number of words and keywords, id. Only the whole text was obtained with the web scraping method and included in the dataset. Name, surname, gender, age and occupation of the person were obtained on this text by using natural language processing methods. The articles that did not mention a single person like articles about Apollo11 team were removed from the dataset since they were not suitable for our study in terms of gender and professional extraction.

When the titles of articles were examined, it was seen that the title included the full name of the person and the age of death. NER (Honnibal and Montani 2017) defines assets such as person, organization, place names and time by using the information available in the texts. Using this method, personal information was obtained from article titles. In order to find the gender of the person, the gendered pronouns were used.

Biographies were collected from Wikipedia using the names of individuals to verify their professional knowledge. “SOC 2018” and “O*NET” datasets containing the occupations and their definitions were used to analyze the occupations over the text (Bureau of Labor Statistics). SOC clearly shows the hierarchy for each occupation and occupational group through major and minor groups. O*NET database was used because it is created based on SOC and also includes alternate titles for each job title. Alternative titles were searched in the Wikipedia biography of the person and a job assignment was made in case of a match. After performing key search, the occupations obtained were assigned to their major groups in order to gather under the group they belong to and to decrease the number of classes for classification process. For instance, as shown in Table 4.1, person’s occupation is operating engineer and according to hierarchical structure of coding its major group is “Construction & Extraction Occupations”. All major groups can be found in Appendix A.

[3] Collected data in form of a dataset can be found here: https://drive.google.com/file/d/1sf6GbdE9dwltAjE5XbOQEdh8jBfXZbcg/view?usp=sharing

Table 4.1. Coding Structure

Coding Structure		
Major Group	47-0000	Construction & Extraction Occupations
Minor Group	47-2000	Construction Trade Workers
Broad Occupation	47-2070	Construction Equiptment Operators
Detailed Occupation	47-2073	Operating Engineers and Other Construction Equipment Operators
O*NET Occupation	47-2073.02	Operating Engineer

4.2. Text Preprocessing

Pre-processing is performed to obtain clean text so that machine learning algorithms can perform better. Pre-processing is task specific. It is an important step to obtain a more accurate result. First, lowercasing was performed. All characters are standardized to prevent error caused by case sensitivity. Lowercasing deals with sparsity issues when your dataset is fairly small. Then, special characters were removed since they are not particularly important in predicting a job title or gender.

Next step was removing stop words. Stop words are a set of commonly used words in a language. The purpose of using stop words is to remove these words from the text and focus on the important words instead of these words. The Python NLTK library is used in most natural language processing applications. In Appendix B, first figure shows NLTK’s stop words set for English. To see the effect of gender pronoun in predicting occupations, two different sets of stop words were formed and used for filtering. As shown in second figure of Appendix B, this set was created by removing gendered pronouns and possessive pronouns from NLTK's stop words. Using this to filter articles, gendered articles were obtained. Since scrubbed articles were obtained by removing gendered pronouns and possessive pronouns from articles, original NLTK's stop words were used for scrubbed articles.

Since words are normally used with a space, any extra space may cause a word to be misinterpreted. Removing white space was performed to clear these extra space. Last step in preprocessing was stemming. Stemming is a text normalization technique and it uses a coarse intuitive process that cuts the ends of words in the hope of correctly transforming words into their root form. For example, the root word of *work* is representative of *work, works, and working*. The root is part of the word you add affixes such as "-ed, -ize, -s, -de, mis". So sometimes taking a word from the root may lead to words that are not real words.

4.3. Feature Extraction

Bag-of-words is an approach used in NLP to represent a text as the multi-set of words (n-grams) that appear in it. This creates a simplified representation (e.g. feature vector) of the text. It is important to remember that bag-of-words does not care about the word order. N-gram extraction was done with TF-IDF techniques which is used for information retrieval to represent how important a specific word or phrase is to a given document. Detailed explanation of TF-IDF was made in the background section.

4.4. Classifiers

4.4.1. SVM

SVM represents the decision boundary using a subset of the training examples, known as support vectors. It tries to find a hyperplane that separates classes from each other and maximizes the margin. SVM is designed for binary classification. For multiclass classification it needs to be extended. Since occupational classification is a multiclass classification problem, two different strategies for SVM will be examined in this section.

One-vs-one (OVO) strategy is not a specific feature of SVM. The purpose of this method is to develop an expert binary classifier for each possible class pair and build an ensemble. The term expert is used because each binary classifier is only trained with two classes, and its function is only to derive the decision boundary between these two classes. Multiclass problem has N classes, the OVO ensemble will be composed by $\frac{N*(N-1)}{2}$. The label assignment is carried out by majority vote. For instance, assume that there are three classes, 0, 1 and 2. The OVO ensemble will be composed of three binary classifiers. The first will discriminante 0 from 1, the second 0 from 2, and the third 1 from 2. When the data point is to be classified, the data point is presented to each binary classifier of the ensemble to create a vector of each classification. The final step of labeling is majority voting.

One-vs-rest (OVR) or one-vs-all (OVA) is another strategy for multiclass classification for SVM. One classifier is used for each class and N classifiers are used in total. For each class it will assume the label of that class as positive and the rest as negative.

4.4.2. Hierarchical Attention Networks (HAN)

HAN is a deep neural network which focuses on the idea that not every word in the sentence is equally important to capture meaning. So is every sentence in the document. For this purpose, it has an architecture with an attention mechanism. The model includes encoder and attention mechanisms. Attention mechanism computes the importance weights of contexts. The same algorithm is applied on word level and sentence level.

$$x_{it} = W_e w_{it} \quad , t \in [1, T] \qquad (4.1)$$

Structured tokens w_{it} represent word i per sentence t and embedding layer assigns multidimentional vectors to each token $W_e w_{it}$. As a result, the projection of a word in a continuous vector space is calculated.

These vectorized tokens are the inputs for word encoder layer.

$$\overrightarrow{h_{it}} = \overrightarrow{GRU}(x_{it}) \, , t \in [1, T] \qquad (4.2)$$

$$\overleftarrow{h_{it}} = \overleftarrow{GRU}(x_{it}), t \in [T, 1] \qquad (4.3)$$

$$h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}] \qquad (4.4)$$

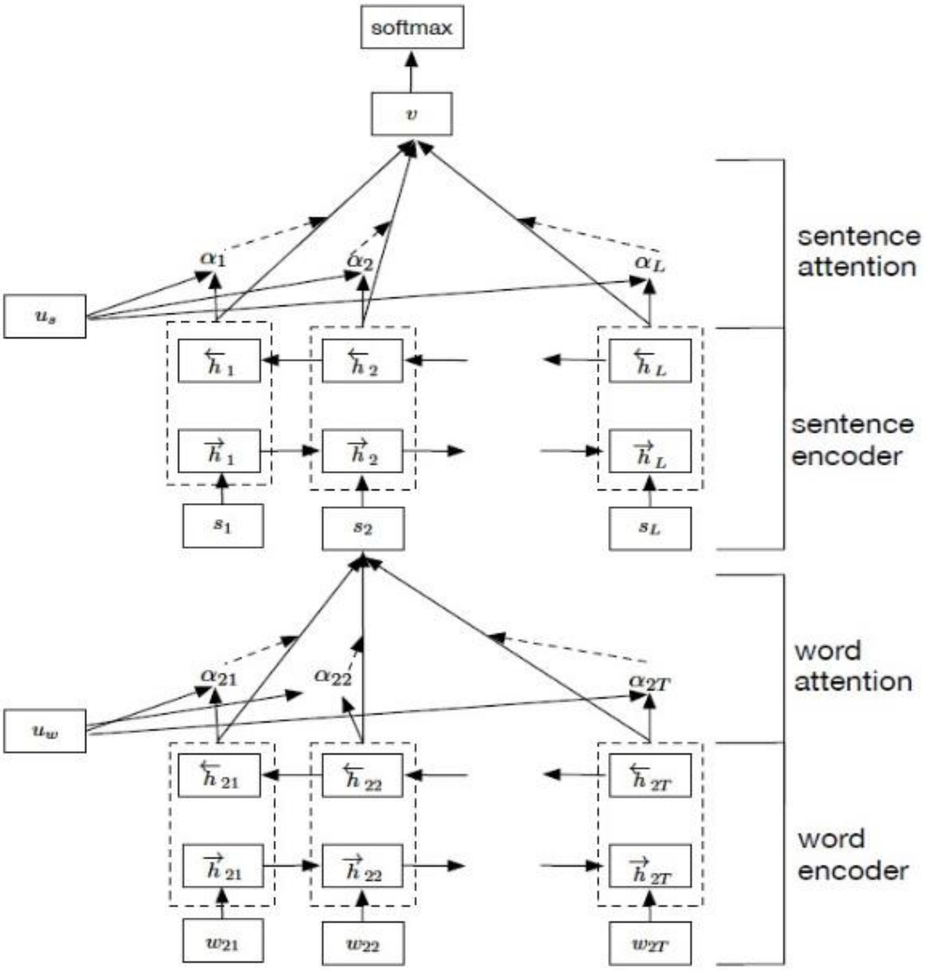


Figure 4.1. Illustration of HAN Architecture

(Source: *Yang, et al. 2016*)

The purpose of this layer is to extract the relevant contexts of each sentence. These contexts per word are called annotations. To do that a bidirectional GRU is applied to get annotations of words by summarizing information from forward and backward directions.

Word attention layer's purpose is to let the model learn through training with randomly initialized weights and biases.

$$u_{it} = tanh(W_w h_{it} + b_w) \qquad (4.5)$$

$$a_{jt} = \frac{exp(u_{it}^T u_w)}{\Sigma_t \, exp(u_{it}^T u_w)} \qquad (4.6)$$

$$s_i = \sum_t a_{it} h_{it} \qquad (4.7)$$

The sum of these importance weights is concatenated with the previously calculated h_{it} called sentence vector s_i. Then the network is run on sentence level with the same procedure as on word level.

4.4.3. DistilBERT

DistilBERT (Sanh, et al. 2019) is a smaller language model from the supervision of BERT (Devlin, et al. 2019). Distillation is a technique used to compress a large model called a teacher into a smaller model called a student. A small model is trained to reproduce the behavior of a larger model. DistilBERT network architecture is a Transformer (Vaswani, et al. 2017) encoder model and it consists of 6-layer, 768-hidden units, 12-heads and 66M parameters. It has half the number of layers compared to BERT, while keeping the hidden representation dimension the same.

4.4.4. Multi-task Learning (MTL)

Multitask learning activities are mainly focused on feature-based and parameter-based approaches. In feature based, it is aimed to learn common attributes to be shared among different tasks. In parameter-based learning, it is aimed to improve the trainings by applying the optimized model coefficients used by one task to other tasks. In feature-based multitasking learning, there are two basic approaches: feature learning approach and deep learning approach. The common features learned are a subset of the original features or are obtained by transferring the original features. The deep learning approach learns common representations with deep architectures with multiple tasks. To do that it uses soft or hard parameter sharing of hidden layers.

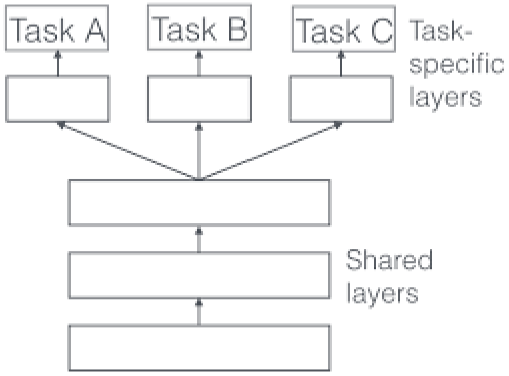


Figure 4.2. Hard Parameter Sharing
(Source: *Ruder 2017*)

In this thesis, hard parameter sharing whose architecture is shown in the Figure 4.2 is used for multitask learning model. Hard parameter sharing approach commonly used in MTL since it greatly reduces the risk of overfitting. The more learned at the same time, the model has to find a representation that captures all the tasks. That causes less chance the overfitting on original task. A separate loss is computed per tasks and they combined into the general loss of the network by weighting each of them. Additionally, MTL allows having a single model for each task rather than a separate model. This helps reduce storage space, decreases training times and is easier to deploy and maintain. Thus storage space reduces and training times decreases.

4.5. Basis of Evaluation

There are several metrics used to evaluate the performance of models. In some cases, different metrics are needed, even if the most used is accuracy. In this section, the metrics that are widely used in the literature and which we use to evaluate our models are discussed.

True positive (TP) is a correctly predicted positive value which means actual class and predicted class are the same. True negative (TN) is a predicted negative value where the actual and predicted class are the same as the actual class is “no” and the predicted class is also “no”. False positive (FP) and false negative (FN) occur when these values are contrary to actual class.

Accuracy is one of the performance measures and basically it is the ratio of correctly predicted observations to total number of observations as in equation 4.8.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.8}$$

Accuracy can be misleading especially for the models that are trained with an imbalanced dataset. In a problem with a large class imbalance, a model can predict the value of the majority class for all predictions and achieve a high classification accuracy but this is not a successful model. Alternative metrics should be used in such cases for evaluation. Precision measures how much of the positive definitions are really correct. It is the ratio between the correct predictions and the total predictions as in equation 4.9. Recall is the metric that measures how much of the true positives are correctly defined by taking the ratio of correct predictions and total number of correct items as in equation 4.10.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.10}$$

From equation 4.9 and 4.10, we can calculate F1 measure which is a weighted average of precision and recall as in equation 4.11.

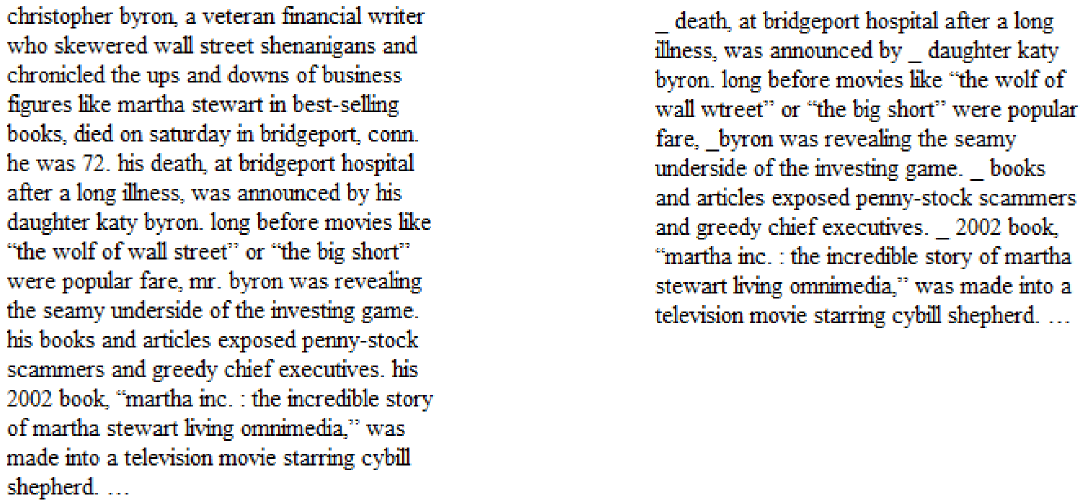
$$F1 = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{4.11}$$

CHAPTER 5

EXPERIMENTAL RESULTS

As explained in the previous sections, a new dataset and different models were used in this thesis in order to detect gender bias. Since the variables of gender and profession are independent from each other, it is expected that the results we obtain from the models in the occupation estimation of the people will be the same. In this section, firstly the results of SVM model, then the results of HAN and DistilBERT model, and finally the results obtained with MTL will be examined. True positive rates calculated separately for each gender, the performance of models, the effect of first sentence on the prediction will be discussed. Since gender and occupation variables are independent from each other, the occupational information that models predict for each gender should be the same with given articles and scrubbed articles as input. To evaluate that, true positive rates of each gender will be used as basis. If there is a difference between these two results, we can say that gender information plays a role in the estimation, there is a dependency between the two variables, and therefore gender bias occurs. Figure 5.1 shows an example of article with gender indicators and scrubbed version of the article. These are the two versions of articles given to the models. Scrubbed version does not include first sentence and gender indicators.

The reason for the deletion of the first sentence from the article is that the person's name and occupational knowledge are included in the estimation and contain the gender information due to the nature of the name. According to the scrubbing method, gender indicators should be removed from the article. In order to investigate the effect of the first sentence on the prediction, three situations -sample article, scrubbed version of it and scrubbed version with the first sentence- were examined with the DistilBERT model. It was examined in detail in the DistilBERT section.



christopher byron, a veteran financial writer who skewered wall street shenanigans and chronicled the ups and downs of business figures like martha stewart in best-selling books, died on saturday in bridgeport, conn. he was 72. his death, at bridgeport hospital after a long illness, was announced by his daughter katy byron. long before movies like "the wolf of wall street" or "the big short" were popular fare, mr. byron was revealing the seamy underside of the investing game. his books and articles exposed penny-stock scammers and greedy chief executives. his 2002 book, "martha inc. : the incredible story of martha stewart living omnimedia," was made into a television movie starring cybill shepherd. …

_ death, at bridgeport hospital after a long illness, was announced by _ daughter katy byron. long before movies like "the wolf of wall wtreet" or "the big short" were popular fare, _byron was revealing the seamy underside of the investing game. _ books and articles exposed penny-stock scammers and greedy chief executives. _ 2002 book, "martha inc. : the incredible story of martha stewart living omnimedia," was made into a television movie starring cybill shepherd. …

Figure 5.1. Sample Article and its Scrubbed Version

5.1. SVM

Word unigrams, bigrams and trigrams are used in SVM model with the following parameters for the feature sets:

- TF-IDF weighting
- Minimum document frequency is set as 2
- Maximum document frequency is set as 1.0 which means terms that ocur in all documents will be ignored.

Dimensionality reduction was done using the TruncatedSVD function from the scikit-learn library (Pedregosa, et al. 2011). Number of dimensions is set to 300 in the final model. Before examining the results, let's first look at the distribution of major titles in Figure 5.2 and gender distribution of major titles in Table 5.1. We see in here the dataset is not balanced.

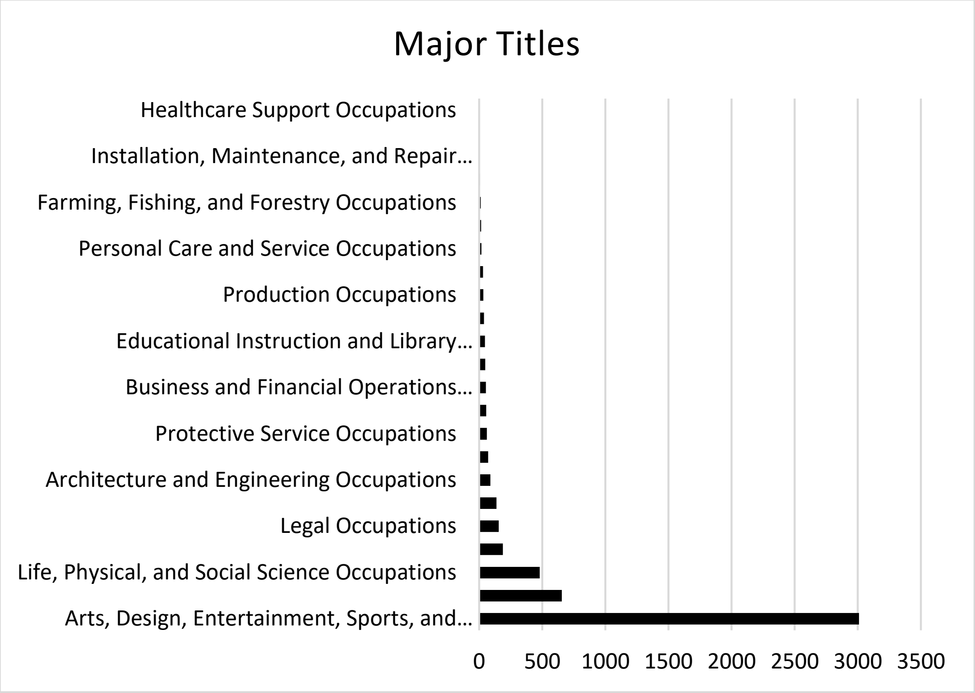


Figure 5.2. Distribution of Major Titles

Table 5.1. Gender Distribution of Major Titles

Major Titles	Male	Female
Arts, Design, Entertaintment, Sports, and Media Occupations	2211	799
Management Occupations	555	99
Life, Physical, and Social Science Occupations	393	87
Community and Social Service Occupations	150	37
Legal Occupations	120	34
Healthcare Practitioners and Technical Occupations	108	28
Architecture and Engineering Occupations	82	8
Military Specific Occupations	67	4
Protective Service Occupations	55	7
Computer and Mathematical Occupations	47	10
Business and Financial Operations Occupations	40	13
Food Preparation and Serving Related Occupations	33	16
Educational Instruction and Library Occupations	30	25
Transportation and Metarial Moving Occupations	29	15
Production Occupations	23	3
Sales and Related Occupations	22	1
Personal Care and Service Occupations	11	8
Office and Administrative Support Occupations	10	10
Farming, Fishing, and Forestry Occupations	6	2

(Cont.on next page)

Table 5.1. (Cont.)

Construction and Extraction Occupations	5	0
Installation, Maintenance, and Repair Occupations	3	1
Building and Ground Cleaning and Maintenance Occupations	2	0
Healthcare Support Occupations	0	1

In Table 5.1, we see gender distribution of major titles. Since we try to reveal that there is a gender bias due to the dependence of gender and occupation variables and the model reflects the bias that is available, X^2 test was applied using the values in this table. The null hypothesis for this test is that there is no relationship between gender and occupation. Chi-squared statistic is 124.82 and degree of freedom is 22. The obtained p value is 2.3038471882453996e-16. This means a highly statistically significant result. Since the p-value is smaller than 0.01, we can reject the null hypothesis. In other words, there is a dependency.

As seen in Figure 5.2, the dataset is not balanced, so there is an imbalance problem. To solve class imbalance problem, “class_weight” parameter of SVM is set as balanced. OVO strategy was used because we were working on multi-class classification and also micro- average method is used for calculating F1-Score. We evaluated the accuracy of SVM using 10-fold cross-validation. The dataset is divided into 80% training set and 20% test set. Results of the model are given in Table 5.2 and Table 5.3.

Table 5.2. Accuracy and F1 scores of SVM for Occupation Classification.

	CROSS VALIDATION	**TEST SET**	
	Mean Accuracy	Accuracy	F1– Score$_{\mathbf{micro\_avg}}$
ARTICLES	0,67	0,77	0,76
SCRUBBED ARTICLES	0,65	0,75	0,74

TPR_{gap} value is obtained by taking the difference of TPR_{male} and TPR_{female} values. These results obtained for article and scrubbed article should be the same if there is no gender bias but as can be seen from Table 5.3, gender gap between TPR_{male} and TPR_{female} decreased for scrubbed articles. This means there is a gender bias since gender indicators affect the results. In following tables, we see detailed classification report.

Table 5.3. TPRs of Articles and Scrubbed Articles from SVM for Occupation Classification.

	ARTICLES	**SCRUBBED ARTICLES**
$\text{TPR}_{\text{female}}$	0,46261	0,45421
TPR_{male}	0,48592	0,46308
TPR_{gap}	**0,02330**	**0,00886**

Table 5.4. Classification Report of SVM for Articles with Gender Indicators

Major Titles	**Precision**	**Recall**	**F1-Score**
Arts, Design, Entertainment, Sports, and Media Occupations	0.93	0.89	0.91
Management Occupations	0.56	0.52	0.54
Life, Physical, and Social Science Occupations	0.75	0.67	0.71
Community and Social Service Occupations	0.39	0.71	0.50
Legal Occupations	0.82	0.73	0.77
Healthcare Practitioners and Technical Occupations	0.81	0.89	0.85
Architecture and Engineering Occupations	0.64	0.67	0.65
Computer and Mathematical Occupations	0.80	0.47	0.59
Protective Service Occupations	0.50	0.71	0.59
Military Specific Occupations	0.43	0.60	0.50
Business and Financial Operations Occupations	0.25	0.25	0.25
Educational Instruction and Library Occupations	0.23	0.27	0.25
Food Preparation and Serving Related Occupations	0.71	0.83	0.77
Transportation and Material Moving Occupations	0.50	0.50	0.50
Production Occupations	0.33	0.29	0.31
Sales and Related Occupations	0.11	0.50	0.18
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Healthcare Support Occupations	0.00	0.00	0.00
Macro Avg	0.42	0.45	0.42
Weighted Avg	0.79	0.77	0.77

Table 5.5. Classification Report of SVM for Scrubbed Articles

Major Titles	Precision	Recall	F1-Score
Arts, Design, Entertainment, Sports, and Media Occupations	0.92	0.88	0.90
Management Occupations	0.54	0.46	0.50
Life, Physical, and Social Science Occupations	0.74	0.62	0.68
Community and Social Service Occupations	0.37	0.71	0.49
Legal Occupations	0.79	0.73	0.76
Healthcare Practitioners and Technical Occupations	0.81	0.89	0.85
Architecture and Engineering Occupations	0.61	0.67	0.64
Computer and Mathematical Occupations	0.80	0.47	0.59
Protective Service Occupations	0.33	0.43	0.38
Military Specific Occupations	0.41	0.70	0.52
Business and Financial Operations Occupations	0.21	0.25	0.23
Educational Instruction and Library Occupations	0.12	0.18	0.14
Food Preparation and Serving Related Occupations	0.71	0.83	0.77
Transportation and Material Moving Occupations	0.57	0.50	0.53
Production Occupations	0.33	0.29	0.31
Sales and Related Occupations	0.07	0.25	0.11
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.14	0.33	0.20
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Healthcare Support Occupations	0.00	0.00	0.00
Macro Avg	0.40	0.44	0.41
Weighted Avg	0.78	0.75	0.76

5.2. HAN

Neither every sentence in the document nor every word in the sentence is of equal importance. Based on this idea, HAN performs better on prediction. After preprocessing, sentences were tokenized and tokens are vectorized with GloVe's (Pennington, Socher and Manning 2014) pre-trained embeddings in 100 dimensions. Articles and Scrubbed Articles were treated as two separate corpora. Maximum number of sentences is 279 and maximum number of words for sentences is 127 for Articles Corpus. Maximum number of sentences is 278 and maximum number of words for sentences is 119 for Scrubbed Articles Corpus. These values are used in padding to ensure that the input is of equal length. The dataset is divided into 60% training set, %20 validation set, and 20% test set. The model was trained with a number of 7 epochs and batch size of 50. Results of the model are given in Table 5.6 and Table 5.7.

Table 5.6. Accuracy and F1 Scores of HAN for Occupation Classification

	TEST SET	
	Accuracy	F1– Score$_{\textbf{micro\_avg}}$
ARTICLES	0,79	0,79
SCRUBBED ARTICLES	0,76	0,75

Table 5.7. TPRs of Articles and Scrubbed Articles from HAN for Occupation Classification.

	ARTICLES	**SCRUBBED ARTICLES**
TPR$_{female}$	0,32235	0,19980
TPR$_{male}$	0,26135	0,18791
TPR$_{gap}$	**0,06100**	**0,01188**

As seen in SVM, the TPR results obtained for HAN also show differences. These results show that the model also pays attention to gender indicators. Therefore, gender bias occurs. In Table 5.8 and Table 5.9, we see all major titles and their detailed classification report.

Table 5.8. Classification Report of HAN for Articles with Gender Indicators

Major Titles	**Precision**	**Recall**	**F1-Score**
Architecture and Engineering Occupations	0.59	0.62	0.61
Arts, Design, Entertainment, Sports, and Media Occupations	0.92	0.95	0.93

(Cont.on next page)

Table 5.8. (Cont.)

Business and Financial Operations Occupations	0.00	0.00	0.00
Community and Social Service Occupations	0.39	0.32	0.35
Computer and Mathematical Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Educational Instruction and Library Occupations	0.00	0.00	0.00
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Food Preparation and Serving Related Occupations	0.00	0.00	0.00
Healthcare Practitioners and Technical Occupations	0.64	0.82	0.72
Healthcare Support Occupations	0.00	0.00	0.00
Legal Occupations	0.71	0.74	0.72
Life, Physical, and Social Science Occupations	0.76	0.77	0.77
Management Occupations	0.52	0.83	0.64
Military Specific Occupations	0.75	0.18	0.29
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Production Occupations	0.00	0.00	0.00
Protective Service Occupations	0.00	0.00	0.00
Sales and Related Occupations	0.00	0.00	0.00
Transportation and Material Moving Occupations	0.00	0.00	0.00
Macro Avg	0.25	0.25	0.24
Weighted Avg	0.75	0.79	0.76

Table 5.9. Classification Report of HAN for Scrubbed Articles

Major Titles	**Precision**	**Recall**	**F1-Score**
Architecture and Engineering Occupations	0.00	0.00	0.00
Arts, Design, Entertainment, Sports, and Media Occupations	0.83	0.98	0.90
Business and Financial Operations Occupations	0.00	0.00	0.00
Community and Social Service Occupations	0.33	0.03	0.05
Computer and Mathematical Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Educational Instruction and Library Occupations	0.00	0.00	0.00

(Cont.on next page)

Table 5.9. (Cont.)

Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Food Preparation and Serving Related Occupations	0.00	0.00	0.00
Healthcare Practitioners and Technical Occupations	0.63	0.76	0.69
Healthcare Support Occupations	0.00	0.00	0.00
Legal Occupations	1.00	0.38	0.55
Life, Physical, and Social Science Occupations	0.65	0.75	0.70
Management Occupations	0.51	0.68	0.58
Military Specific Occupations	1.00	0.06	0.11
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Production Occupations	0.00	0.00	0.00
Protective Service Occupations	0.00	0.00	0.00
Sales and Related Occupations	0.00	0.00	0.00
Transportation and Material Moving Occupations	0.00	0.00	0.00
Macro Avg	0.24	0.17	0.17
Weighted Avg	0.68	0.76	0.70

5.3. DistilBERT

BERT expects input data in a certain format, with special tokens for the beginning "[CLS]" and to separation or end of sentences "[SEP]". As seen in the Figure 5.3, there is an example of input data with input tokens, input ids and attention mask expected by the BERT model (Wolf, et al. 2019). BERT model also uses segment ids if there is a separation in the input data, these ids show which words belongs to which part. DistilBERT is a distilled version of BERT. As you can see in Figure 5.4, it uses the same input format as BERT but DistilBERT model (Maiya 2020) does not have segment ids option. Using [SEP] tag, you can separate your section but segment ids will be the same (zero) for all words.

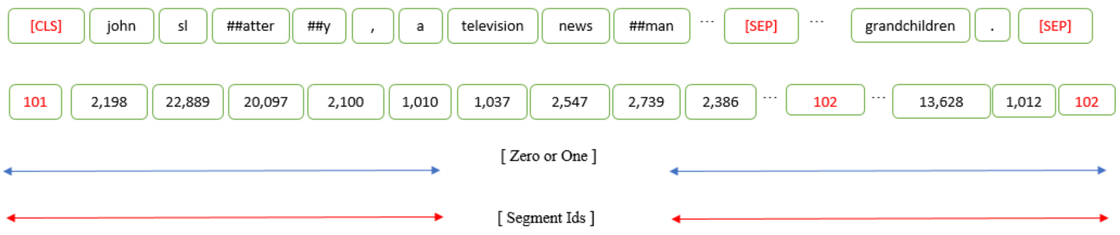


Figure 5.3. Input Format of BERT.

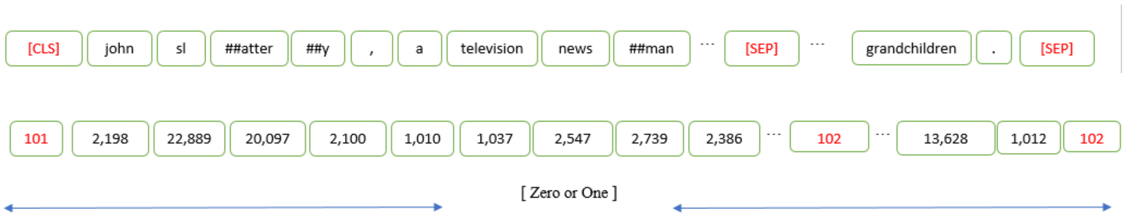


Figure 5.4. Input Format of DistilBERT.

The dataset is divided into 60% training set, %20 validation set and 20% test set. Learning rate is set as 5e-5, epoch number is 3, batch size is 50. Results of the model are given in Table 5.10 and Table 5.11.

Table 5.10. Accuracy and F1 Scores of DistilBERT for Occupation Classification

	TEST SET	
	Accuracy	F1– Score$_{\mathbf{micro\_avg}}$
ARTICLES	0,83	0,83
SCRUBBED ARTICLES	0,81	0,80

Table 5.11. TPRs of Articles and Scrubbed Articles from DistilBERT for Occupation Classification.

	ARTICLES	**SCRUBBED ARTICLES**
TPR_{female}	0,38741	0,33819
TPR_{male}	0,35806	0,33205
TPR_{gap}	**0,02934**	**0,00614**

Table 5.6 shows gender gap is lower for scrubbed articles than articles. The fact that the gender gap decreases in the absence of gender indicators indicates that there is a gender bias.

It was stated that gender bias was observed based on the results obtained so far. The words participating in the prediction are expressed visually from here on. Original output images of DistilBERT highlighting the words contributing to the final label in green, and the detracting words in red, can be found in Appendix C. Representative visuals based on original images will be examined in following figures. Let's see this on the sample article about Christopher Byron. He is a writer and major title of his occupation is “Arts, Design, entertaintment, Sports and Media Occupations”.

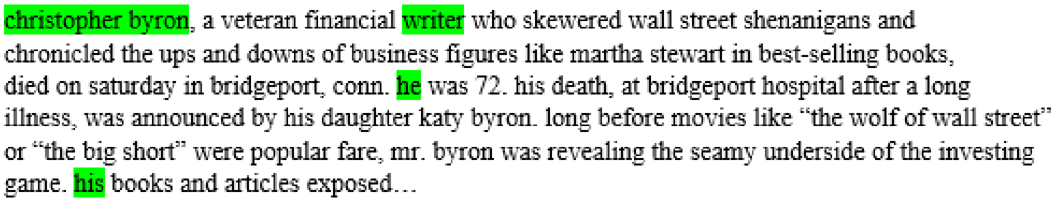

christopher byron, a veteran financial writer who skewered wall street shenanigans and
chronicled the ups and downs of business figures like martha stewart in best-selling books,
died on saturday in bridgeport, conn. he was 72. his death, at bridgeport hospital after a long
illness, was announced by his daughter katy byron. long before movies like “the wolf of wall street”
or “the big short” were popular fare, mr. byron was revealing the seamy underside of the investing
game. his books and articles exposed…

Figure 5.5. Visualization of the Weight of Words for the Sample Article.

In Figure 5.5, gender indicators like “he” or “his” are highlighted in green since they have an impact on predicting major title “Arts, Design, entertaintment, Sports and Media Occupations. Gender gap is shown in Articles part of Table 5.6.

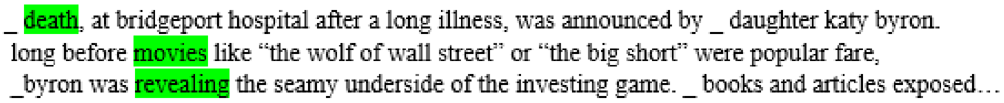

_ death, at bridgeport hospital after a long illness, was announced by _ daughter katy byron.
long before movies like “the wolf of wall street” or “the big short” were popular fare,
_byron was revealing the seamy underside of the investing game. _ books and articles exposed…

Figure 5.6 Visualization of the Weight of Words for Scrubbed Article.

In Figure 5.6, words such as “movies, book, television” are highlighted in green. Occupation prediction is done without gender indicators and first sentence. Predicted major title is “Arts, Design, entertaintment, Sports and Media Occupations”. Figure 5.5 and Figure 5.6 clearly show that gender indicators are included in the estimate when it is in the article. Gender gap is decreased as seen in Scrubbed Articles part of Table 5.6 when gender indicators are scrubbed.

To see the effect of the first sentence on the prediction, the text was separated into the first sentence and the rest by using the [SEP] tag.

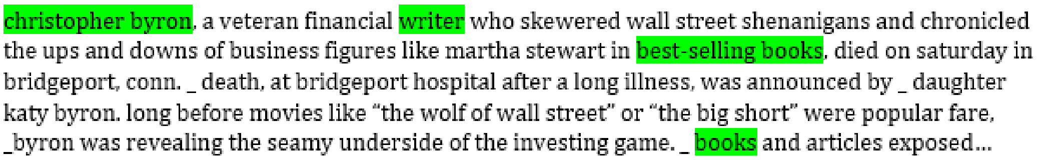

christopher byron, a veteran financial writer who skewered wall street shenanigans and chronicled the ups and downs of business figures like martha stewart in best-selling books, died on saturday in bridgeport, conn. _ death, at bridgeport hospital after a long illness, was announced by _ daughter katy byron. long before movies like “the wolf of wall street” or “the big short” were popular fare, _byron was revealing the seamy underside of the investing game. _ books and articles exposed…

Figure 5.7. Visualization of the Weight of Words for Scrubbed Article with its First Line.

In Figure 5.7, the words highlighted by green are the ones that are considered important in prediction. First name of person and occupation information are highlighted in green. Predicted major title is “Arts, Design, entertaintment, Sports and Media Occupations”. Since the person's name has gender information by nature, its inclusion in the estimation causes gender bias. There is an increase in the gender gap obtained from the article which has no gender indicator other than the name. Table 5.12 shows that gender gap has increased.

Table 5.12 TPRs of Articles and Scrubbed Articles with First Lines from DistilBERT for Occupation Classification.

	SCRUBBED ARTİCLES WİTH FİRST LİNE
TPR$_{female}$	0,35345
TPR$_{male}$	0,32357
TPR$_{gap}$	**0,02987**

In Table 5.13 and Table 5.14, we see all major titles and their detailed classification report for DistilBERT.

Table 5.13 Classification Report of DistilBERT for Articles with Gender Indicators

Major Titles	**Precision**	**Recall**	**F1-Score**
Architecture and Engineering Occupations	0.90	0.56	0.69
Arts, Design, Entertainment, Sports, and Media Occupations	0.92	0.96	0.94
Business and Financial Operations Occupations	0.00	0.00	0.00
Community and Social Service Occupations	0.67	0.43	0.52

(Cont.on next page)

Table 5.13. (Cont.)

Computer and Mathematical Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Educational Instruction and Library Occupations	0.00	0.00	0.00
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Food Preparation and Serving Related Occupations	1.00	0.23	0.38
Healthcare Practitioners and Technical Occupations	0.82	0.68	0.74
Healthcare Support Occupations	0.00	0.00	0.00
Legal Occupations	0.86	0.91	0.89
Life, Physical, and Social Science Occupations	0.67	0.89	0.76
Management Occupations	0.61	0.82	0.70
Military Specific Occupations	0.81	0.76	0.79
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Production Occupations	0.00	0.00	0.00
Protective Service Occupations	0.75	0.43	0.55
Sales and Related Occupations	0.00	0.00	0.00
Transportation and Material Moving Occupations	0.67	0.50	0.57
Macro Avg	0.41	0.34	0.36
Weighted Avg	0.79	0.83	0.80

Table 5.14. Classification Report of DistilBERT for Scrubbed Articles

Major Titles	**Precision**	**Recall**	**F1-Score**
Architecture and Engineering Occupations	1.00	0.25	0.40
Arts, Design, Entertainment, Sports, and Media Occupations	0.89	0.96	0.93
Business and Financial Operations Occupations	0.00	0.00	0.00
Community and Social Service Occupations	0.85	0.30	0.44
Computer and Mathematical Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Educational Instruction and Library Occupations	0.00	0.00	0.00
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00

(Cont.on next page)

Table 5.14. (Cont.)

Food Preparation and Serving Related Occupations	0.60	0.23	0.33
Healthcare Practitioners and Technical Occupations	0.88	0.68	0.77
Healthcare Support Occupations	0.00	0.00	0.00
Legal Occupations	0.84	0.76	0.80
Life, Physical, and Social Science Occupations	0.68	0.77	0.72
Management Occupations	0.55	0.81	0.66
Military Specific Occupations	0.70	0.94	0.80
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Production Occupations	0.00	0.00	0.00
Protective Service Occupations	1.00	0.43	0.60
Sales and Related Occupations	0.00	0.00	0.00
Transportation and Material Moving Occupations	0.67	0.50	0.57
Macro Avg	0.41	0.32	0.33
Weighted Avg	0.78	0.81	0.78

5.4. MTL

Using the MTL model, the results obtained when the gender and occupation are learned together on the articles will be examined in here. The MTL model used is based on the architecture of HAN. Hard parameter sharing approach was used. After shared layers, there is task specific layer to get predictions for gender and occupation. Loss is calculated by binary cross entropy for gender and categorical cross entropy for occupation. There are assigned weights to each task to emphasize that one is more important than the other. In our cases, occupation loss is more weighted than gender loss. Adam optimizer tries optimize this weighted sum instead of the individual losses. The dataset is divided into 60% training set, %20 validation set, and 20% test set. The model was trained with a number of 8 epochs and batch size of 50. Results of the model are given in Table 5.15, Table 5.16 and Table 5.17.

Table 5.15. Accuracy and F1-Score for Occupation Prediction.

	TEST SET	
	Accuracy	F1– Score$_{\text{micro\_avg}}$
ARTICLES	0,77	0,76
SCRUBBED ARTICLES	0,76	0,76

Table 5.16. Accuracy for Gender Prediction.

	TEST SET
	Accuracy
ARTICLES	0,99
SCRUBBED ARTICLES	0,90

Table 5.17. TPRs of Articles and Scrubbed Articles from MTL for Occupation Classification.

	ARTICLES	**SCRUBBED ARTICLES**
$\text{TPR}_{\text{female}}$	0,17429	0,21636
TPR_{male}	0,23031	0,19929
TPR_{gap}	**0,05601**	**0,01706**

Different weight values were tried in the calculation of loss, the best result among the tried weights is obtained when it is "1." for occupation and ".9" for gender. The model performs good as well for scrubbed articles as it does in the articles with gender indicators. As seen in Table 5.17, the TPR results are different from each other for articles and scrubbed articles. Gender gap has decreased for scrubbed articles. This means that here is also gender bias. Learning gender and profession together did not affect the performance of the model to a great extent. In Table 5.18 and Table 5.19, we see all major titles, genders and their detailed classification report of MTL for articles with gender indicators. In Table 5.20 and Table 5.21, we see these detailed classification report for scrubbed articles.

Table 5.18. Classification Report of MTL for Articles with Gender Indicators

Major Titles	Precision	Recall	F1-Score
Architecture and Engineering Occupations	0.50	0.12	0.20
Arts, Design, Entertainment, Sports, and Media Occupations	0.88	0.96	0.92
Business and Financial Operations Occupations	0.00	0.00	0.00
Community and Social Service Occupations	0.73	0.30	0.42
Computer and Mathematical Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Educational Instruction and Library Occupations	0.00	0.00	0.00
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Food Preparation and Serving Related Occupations	0.00	0.00	0.00
Healthcare Practitioners and Technical Occupations	0.67	0.59	0.62
Healthcare Support Occupations	0.00	0.00	0.00
Legal Occupations	0.63	0.79	0.70
Life, Physical, and Social Science Occupations	0.72	0.65	0.68
Management Occupations	0.48	0.82	0.60
Military Specific Occupations	0.00	0.00	0.00
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Production Occupations	0.00	0.00	0.00
Protective Service Occupations	0.00	0.00	0.00
Sales and Related Occupations	0.00	0.00	0.00
Transportation and Material Moving Occupations	0.00	0.00	0.00
Macro Avg	0.22	0.20	0.20
Weighted Avg	0.71	0.77	0.73

Table 5.19. Gender Classification Report of MTL for Articles with Gender Indicators

Gender	Precision	Recall	F1-Score
Female	0.96	1.00	0.98
Male	1.00	0.99	0.99
Macro Avg	0.98	0.99	0.99
Weighted Avg	0.99	0.99	0.99

Table 5.20. Classification Report of MTL for Scrubbed Articles

Major Titles	**Precision**	**Recall**	**F1-Score**
Architecture and Engineering Occupations	0.25	0.12	0.17
Arts, Design, Entertainment, Sports, and Media Occupations	0.89	0.96	0.92
Business and Financial Operations Occupations	0.00	0.00	0.00
Community and Social Service Occupations	0.89	0.22	0.35
Computer and Mathematical Occupations	0.00	0.00	0.00
Construction and Extraction Occupations	0.00	0.00	0.00
Educational Instruction and Library Occupations	0.00	0.00	0.00
Farming, Fishing, and Forestry Occupations	0.00	0.00	0.00
Food Preparation and Serving Related Occupations	0.00	0.00	0.00
Healthcare Practitioners and Technical Occupations	0.76	0.38	0.51
Healthcare Support Occupations	0.00	0.00	0.00
Legal Occupations	0.73	0.65	0.69
Life, Physical, and Social Science Occupations	0.52	0.82	0.63
Management Occupations	0.53	0.74	0.62
Military Specific Occupations	0.00	0.00	0.00
Office and Administrative Support Occupations	0.00	0.00	0.00
Personal Care and Service Occupations	0.00	0.00	0.00
Production Occupations	0.00	0.00	0.00
Protective Service Occupations	0.00	0.00	0.00
Sales and Related Occupations	0.00	0.00	0.00
Transportation and Material Moving Occupations	0.00	0.00	0.00
Macro Avg	0.22	0.19	0.18
Weighted Avg	0.71	0.76	0.72

Table 5.21. Gender Classification Report of SVM for Scrubbed Articles

Gender	**Precision**	**Recall**	**F1-Score**
Female	0.92	0.60	0.73
Male	0.89	0.98	0.94
Macro Avg	0.91	0.79	0.83
Weighted Avg	0.90	0.90	0.89

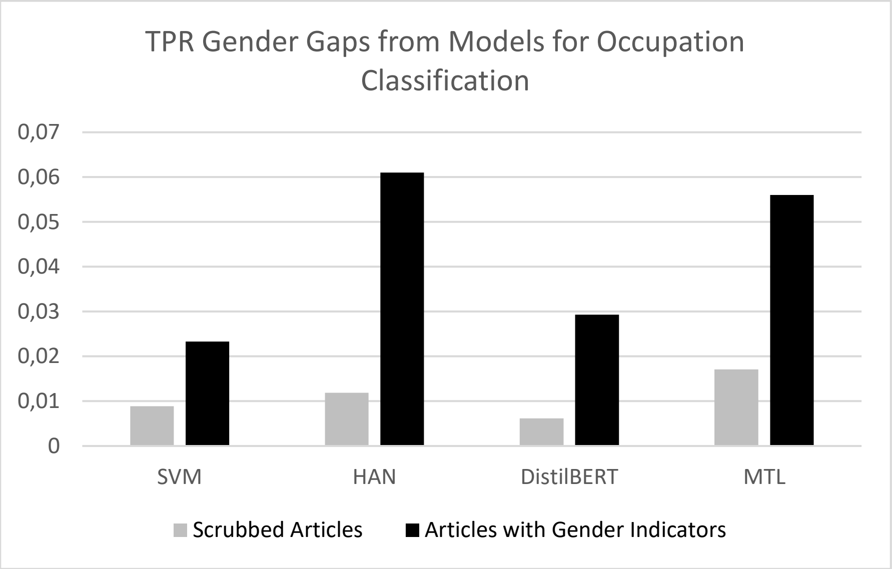


Figure 5.8. TPR Gender Gaps from Models for Occupation Classification.

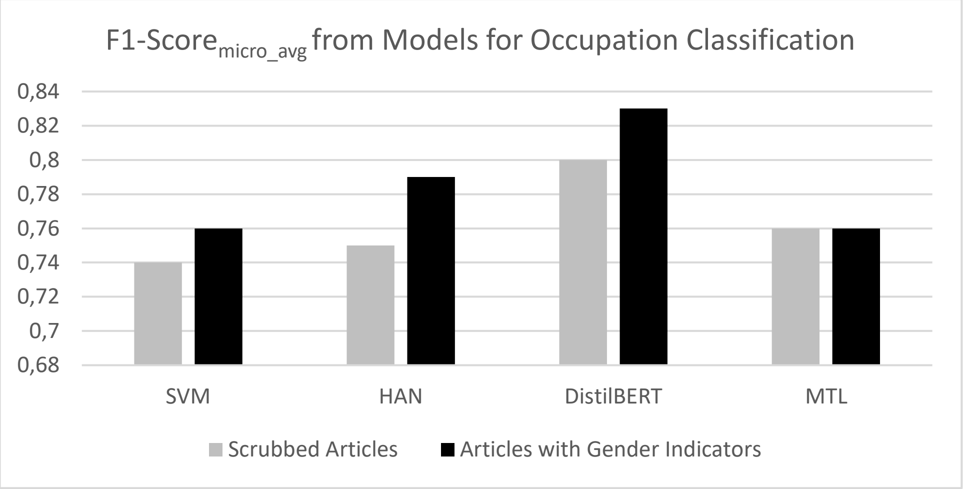


Figure 5.9. F1-Score$_{\text{micro\_avg}}$ from Models for Occupation Classification.

TPR gender gaps obtained from models for occupation estimation are given in Figure 5.8 together. Each model shows there is a gender bias and this situation does not change in the scenario where gender and occupation variables are learned together. As seen in previous models, MTL model showed gender bias when gender indicators were included in the estimation. If there was an serious increase in the gender gap obtained for scrubbed articles from the model where gender and occupation are learned together, it could be said that learning gender and occupation at the same time increases gender bias.

All F1-Score$_{micro\_avg}$ of models can be found together in Figure 5.9. Here we see that of all models, distilbert performed best and SVM the lowest. Classic machine learning models such as SVM do not have a contextual understanding and also do not preserve word order. For this reason, deep learning models that we used perform better. HAN performed better than SVM because it has an attention mechanism to understand which sentence and also which word is important to capture meaning. Our MTL model is based on HAN so it performs close to HAN. DistiBERT is better than HAN because it has multihead attention and also it is a pretrained language model.

CHAPTER 6

CONCLUSIONS

People can develop a bias against anything consciously or unconsciously. These prejudices are reflected in their behavior, speeches and writings. As artificial intelligence is beginning to be included in our lives in many areas, researchers should be aware of the fact that it can reinforce social biases that are already exist in society. In this context, this thesis study tries to detect gender bias, one of the social biases.

We started by creating new dataset with NYT obituaries. In order to reveal gender bias, we used the scrubbing method, which removed the first sentences and gender indicators from the articles. We tested the dataset with different models and examined the TPR gender gaps obtained from occupation estimation to check whether there was gender bias. Within the scope of this thesis, gender bias analysis was made on models that classify only occupation and the model in which gender and occupational variables were learned together. It has been seen in the results obtained from the models used for occupation prediction that the gender gap in the results obtained from the scrubbing method is less than the gender gap in the articles containing the gender indicators. This means that gender indicators play a role in predicting the occupation and this creates gender bias.

Although our dataset is small, we removed articles about group such as Apollo 11 team and other type of data such as media. In any future work, videos can be converted into texts and included in the dataset to collect more data. Another issue that needs to be addressed in any future work is that the articles contain the names of one's relatives, even if they were written about a single person. For classic machine learning models, the gender population that will occur according to the names included in the article may have an impact on the prediction.

We also only focused on gender bias in our thesis work but there are other biases such as race or socioecomic status that may have effect on occupation classification. In the future, the effects of these biases and how they can be reduced without losing information can be explored.

REFERENCES

Bölükbaşı Tolga, Kai-Wei Chang, James Zou, Vankatesh Saligrama, and Adam Kalai. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embedding." Advance in Neural Information Processing Systems. pg: 8.

Buolamwini Joy, and Timnit Gebru. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pg: 7.

Bureau of Labor Statistics. n.d. https://www.bls.gov/soc/2018 (accessed September 20, 2019). pg:11

Caplar Neven, Sandro Tacchella, and Simon Birrer. (2017). "Quantitative Evaluation of Gender Bias in Astronomical Publications from Citation Counts." Nature Astronomy. pg: 8.

Cho Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. (2014). "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint*. pg: 5.

Cortes Corinna, and Vladimir Vapnik. (1995). "Support-Vector Networks." *Machine Learning*. pg: 2.

De-Artega Maria, et al. (2019). "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Settings." *ACM Conference on Fairness, Accountability, and Transparency.* pg: 9.

Devlin Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *ArXiv*.

Fu Liye, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. (2016). "Tie Breaker: Using Language Models to Quantify Gender Bias in Sports Journalism." *The 25th International Joint Conference on Artificial Intelligence.* pg: 9.

Honnibal Matthew, and Ines Montani. (2017). "Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing". pg: 11.

Jurafsky Daniel, and Martin James. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* pg: 3.

Maiya Arun S. (2020). "ktrain: A Low-Code Library for Augmented Machine Learning." *ArXiv*. pg: 28.

Pedregosa Fabian, et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research 12*. pg: 21.

Pennington Jeffrey, Richard Socher, and Christopher Manning. (2014). "Glove: Global Vectors for Word Representation." *EMNLP*. pg: 25.

Ruder Sebestian. (2017). *An Overview of Multi-Task Learning in Deep Neural Networks.* https://ruder.io/multi-task/. pg:17.

Sanh Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *ArXiv*.

Singular Value Decomposition. n.d. https://en.wikipedia.org/wiki/Singular_value_decomposition. (accessed July 23, 2020). pg: 4.

Thelwall Mike. (2018). "Do females create higher impact research? Scopus Citations and Mendeley Readers for Articles from Five Countries." *Journal of Informetrics* (Journal of Informetrics) 12(4). pg: 9.

Thelwall Mike. "Gender Bias in Sentiment Analysis." (Online Information Review) 42, no. 1 (2018): 7.

Turing Alan. (1950). "Computing Machinery and Intelligence." (Mind) LIX, no. 236. pg: 1.

Vaswani Ashish, et al. (2017). "Attention is All you Need." NIPS.

Wagner Claudia, David Garcia, Mohsen Jadidi, and Markus Strohmaier. (2015). "Its's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia." 9th International AAAI Conference on Weblogs and Social Media. pg: 8.

Wolf Thomas, et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." *ArXiv*. pg: 28.

Yang Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. (2016). "Hierarchical Attention Networks for Document Classification." *Association for Computational Linguistics*.

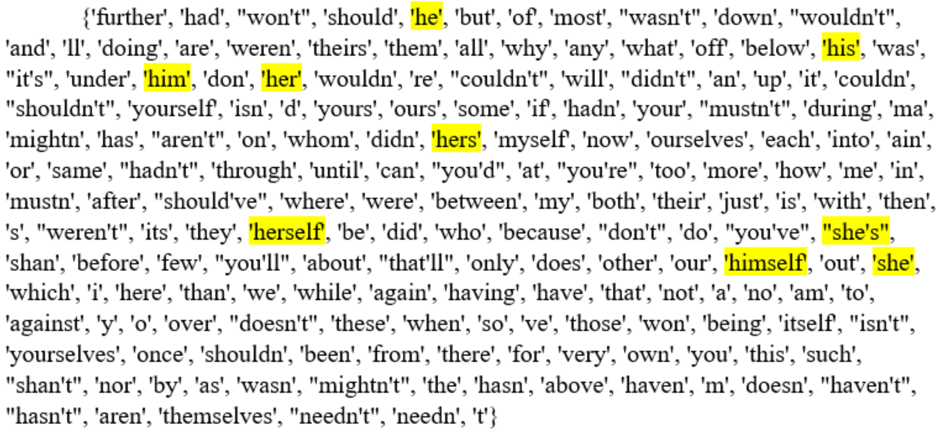
APPENDIX A

Classification Codes (SOC Major Group Code and Title)

Code	***Title***
11-0000	Management Occupations
13-0000	Business and Financial Operations Occupations
15-0000	Computer and Mathematical Occupations
17-0000	Architecture and Engineering Occupations
19-0000	Life, Physical, and Social Science Occupations
21-0000	Community and Social Service Occupations
23-0000	Legal Occupations
25-0000	Education, Training, and Library Occupations
27-0000	Arts, Design, Entertainment, Sports, and Media Occupations
29-0000	Healthcare Practitioners and Technical Occupations
31-0000	Healthcare Support Occupations
33-0000	Protective Service Occupations
35-0000	Food Preparation and Serving Related Occupations
37-0000	Building and Grounds Cleaning and Maintenance Occupations
39-0000	Personal Care and Service Occupations
41-0000	Sales and Related Occupations
43-0000	Office and Administrative Support Occupations
45-0000	Farming, Fishing, and Forestry Occupations
47-0000	Construction and Extraction Occupations
49-0000	Installation, Maintenance, and Repair Occupations
51-0000	Production Occupations
53-0000	Transportation and Material Moving Occupations
55-0000	Military Specific Occupations

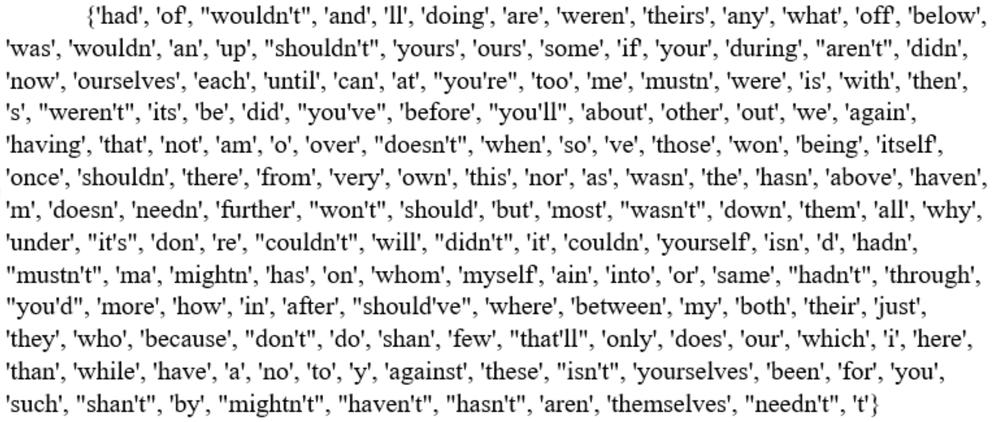
APPENDIX B

NLTK Stop Words Set



```
{'further', 'had', "won't", 'should', 'he', 'but', 'of', 'most', "wasn't", 'down', "wouldn't",
'and', 'll', 'doing', 'are', 'weren', 'theirs', 'them', 'all', 'why', 'any', 'what', 'off', 'below', 'his', 'was',
"it's", 'under', 'him', 'don', 'her', 'wouldn', 're', "couldn't", 'will', "didn't", 'an', 'up', 'it', 'couldn',
"shouldn't", 'yourself', 'isn', 'd', 'yours', 'ours', 'some', 'if', 'hadn', 'your', "mustn't", 'during', 'ma',
'mightn', 'has', "aren't", 'on', 'whom', 'didn', 'hers', 'myself', 'now', 'ourselves', 'each', 'into', 'ain',
'or', 'same', "hadn't", 'through', 'until', 'can', "you'd", 'at', "you're", 'too', 'more', 'how', 'me', 'in',
'mustn', 'after', "should've", 'where', 'were', 'between', 'my', 'both', 'their', 'just', 'is', 'with', 'then',
's', "weren't", 'its', 'they', 'herself', 'be', 'did', 'who', 'because', "don't", 'do', "you've", "she's",
'shan', 'before', 'few', "you'll", 'about', "that'll", 'only', 'does', 'other', 'our', 'himself', 'out', 'she',
'which', 'i', 'here', 'than', 'we', 'while', 'again', 'having', 'have', 'that', 'not', 'a', 'no', 'am', 'to',
'against', 'y', 'o', 'over', "doesn't", 'these', 'when', 'so', 've', 'those', 'won', 'being', 'itself', "isn't",
'yourselves', 'once', 'shouldn', 'been', 'from', 'there', 'for', 'very', 'own', 'you', 'this', 'such',
"shan't", 'nor', 'by', 'as', 'wasn', "mightn't", 'the', 'hasn', 'above', 'haven', 'm', 'doesn', "haven't",
"hasn't", 'aren', 'themselves', "needn't", 'needn', 't'}
```

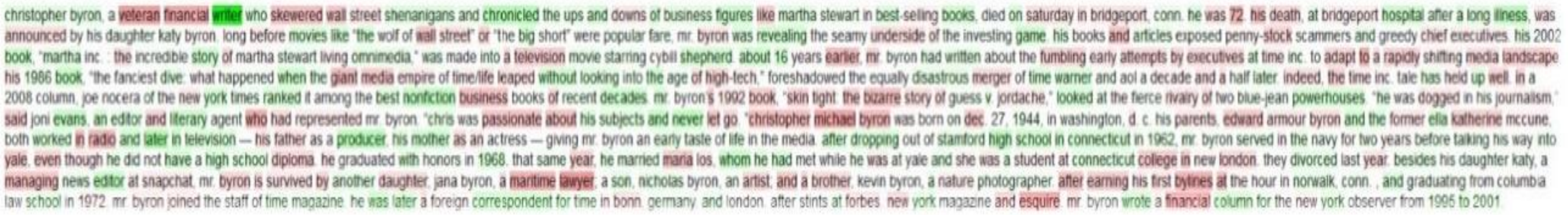
Stop Words without Gender Indicators



```
{'had', 'of', "wouldn't", 'and', 'll', 'doing', 'are', 'weren', 'theirs', 'any', 'what', 'off', 'below',
'was', 'wouldn', 'an', 'up', "shouldn't", 'yours', 'ours', 'some', 'if', 'your', 'during', "aren't", 'didn',
'now', 'ourselves', 'each', 'until', 'can', 'at', "you're", 'too', 'me', 'mustn', 'were', 'is', 'with', 'then',
's', "weren't", 'its', 'be', 'did', "you've", 'before', "you'll", 'about', 'other', 'out', 'we', 'again',
'having', 'that', 'not', 'am', 'o', 'over', "doesn't", 'when', 'so', 've', 'those', 'won', 'being', 'itself',
'once', 'shouldn', 'there', 'from', 'very', 'own', 'this', 'nor', 'as', 'wasn', 'the', 'hasn', 'above', 'haven',
'm', 'doesn', 'needn', 'further', "won't", 'should', 'but', 'most', "wasn't", 'down', 'them', 'all', 'why',
'under', "it's", 'don', 're', "couldn't", 'will', "didn't", 'it', 'couldn', 'yourself', 'isn', 'd', 'hadn',
"mustn't", 'ma', 'mightn', 'has', 'on', 'whom', 'myself', 'ain', 'into', 'or', 'same', "hadn't", 'through',
"you'd", 'more', 'how', 'in', 'after', "should've", 'where', 'between', 'my', 'both', 'their', 'just',
'they', 'who', 'because', "don't", 'do', 'shan', 'few', "that'll", 'only', 'does', 'our', 'which', 'i', 'here',
'than', 'while', 'have', 'a', 'no', 'to', 'y', 'against', 'these', "isn't", 'yourselves', 'been', 'for', 'you',
'such', "shan't", 'by', "mightn't", "haven't", "hasn't", 'aren', 'themselves', "needn't", 't'}
```

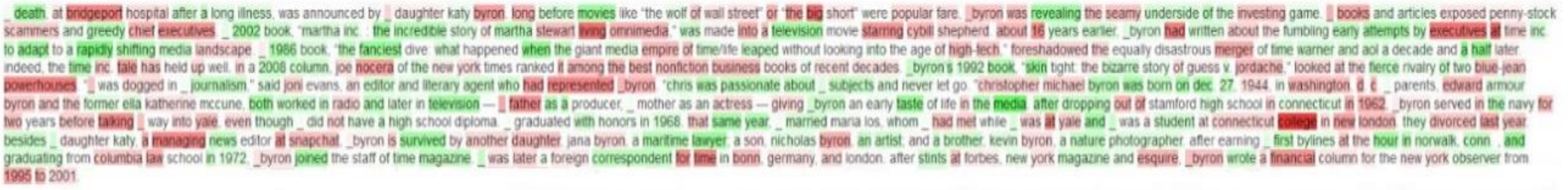
APPENDIX C

Visualization of the Weight of Words for the Sample Article



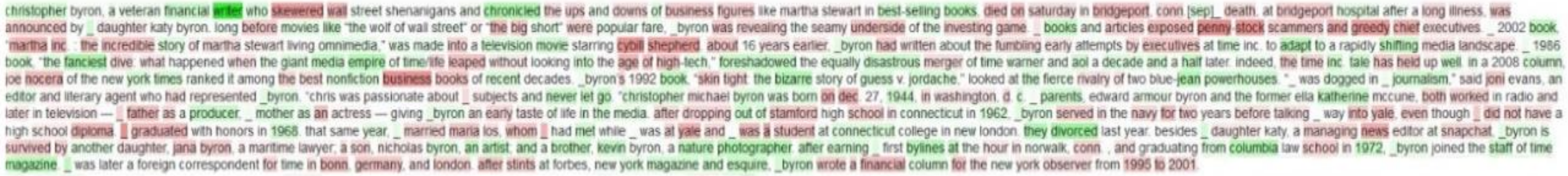
christopher byron, a veteran financial writer who skewered wall street shenanigans and chronicled the ups and downs of business figures like martha stewart in best-selling books, died on saturday in bridgeport, conn. he was 72. his death, at bridgeport hospital after a long illness, was
announced by his daughter katy byron. long before movies like "the wolf of wall street" or "the big short" were popular fare, mr. byron was revealing the seamy underside of the investing game. his books and articles exposed penny-stock scammers and greedy chief executives. his 2002
book, "martha inc. : the incredible story of martha stewart living omnimedia," was made into a television movie starring cybill shepherd. about 16 years earlier, mr. byron had written about the fumbling early attempts by executives at time inc. to adapt to a rapidly shifting media landscape.
his 1986 book, "the fanciest dive: what happened when the giant media empire of time/life leaped without looking into the age of high-tech," foreshadowed the equally disastrous merger of time warner and aol a decade and a half later. indeed, the time inc. tale has held up well. in a
2008 column, joe nocera of the new york times ranked it among the best nonfiction business books of recent decades. mr. byron's 1992 book, "skin tight: the bizarre story of guess v. jordache," looked at the fierce rivalry of two blue-jean powerhouses. "he was dogged in his journalism,"
said joni evans, an editor and literary agent who had represented mr. byron. "chris was passionate about his subjects and never let go. "christopher michael byron was born on dec. 27, 1944, in washington, d. c. his parents, edward armour byron and the former ella katherine mccune,
both worked in radio and later in television — his father as a producer, his mother as an actress — giving mr. byron an early taste of life in the media. after dropping out of stamford high school in connecticut in 1962, mr. byron served in the navy for two years before talking his way into
yale, even though he did not have a high school diploma. he graduated with honors in 1968. that same year, he married maria los, whom he had met while he was at yale and she was a student at connecticut college in new london. they divorced last year. besides his daughter katy, a
managing news editor at snapchat, mr. byron is survived by another daughter, jana byron, a maritime lawyer; a son, nicholas byron, an artist; and a brother, kevin byron, a nature photographer. after earning his first bylines at the hour in norwalk, conn. , and graduating from columbia
law school in 1972. mr. byron joined the staff of time magazine. he was later a foreign correspondent for time in bonn, germany, and london. after stints at forbes, new york magazine and esquire, mr. byron wrote a financial column for the new york observer from 1995 to 2001.

Visualization of the Weight of Words for Scrubbed Article



_ death, at bridgeport hospital after a long illness, was announced by _ daughter katy byron. long before movies like "the wolf of wall street" or "the big short" were popular fare, _byron was revealing the seamy underside of the investing game. _ books and articles exposed penny-stock
scammers and greedy chief executives. _ 2002 book, "martha inc. : the incredible story of martha stewart living omnimedia," was made into a television movie starring cybill shepherd. about 16 years earlier, _byron had written about the fumbling early attempts by executives at time inc.
to adapt to a rapidly shifting media landscape. _ 1986 book, "the fanciest dive: what happened when the giant media empire of time/life leaped without looking into the age of high-tech," foreshadowed the equally disastrous merger of time warner and aol a decade and a half later.
indeed, the time inc. tale has held up well. in a 2008 column, joe nocera of the new york times ranked it among the best nonfiction business books of recent decades. _byron's 1992 book, "skin tight: the bizarre story of guess v. jordache," looked at the fierce rivalry of two blue-jean
powerhouses. "_ was dogged in _ journalism," said joni evans, an editor and literary agent who had represented _byron. "chris was passionate about _ subjects and never let go. "christopher michael byron was born on dec. 27, 1944, in washington, d. c. _ parents, edward armour
byron and the former ella katherine mccune, both worked in radio and later in television — _ father as a producer, _ mother as an actress — giving _byron an early taste of life in the media. after dropping out of stamford high school in connecticut in 1962, _byron served in the navy for
two years before talking _ way into yale, even though _ did not have a high school diploma. _ graduated with honors in 1968. that same year, _ married maria los, whom _ had met while _ was at yale and _ was a student at connecticut college in new london. they divorced last year.
besides _ daughter katy, a managing news editor at snapchat, _byron is survived by another daughter, jana byron, a maritime lawyer; a son, nicholas byron, an artist; and a brother, kevin byron, a nature photographer. after earning _ first bylines at the hour in norwalk, conn. , and
graduating from columbia law school in 1972, _byron joined the staff of time magazine. _ was later a foreign correspondent for time in bonn, germany, and london. after stints at forbes, new york magazine and esquire, _byron wrote a financial column for the new york observer from
1995 to 2001.

Visualization of the Weight of Words for Scrubbed Article with its First Line



christopher byron, a veteran financial writer who skewered wall street shenanigans and chronicled the ups and downs of business figures like martha stewart in best-selling books, died on saturday in bridgeport, conn [sep]_ death, at bridgeport hospital after a long illness, was
announced by _ daughter katy byron. long before movies like "the wolf of wall street" or "the big short" were popular fare, _byron was revealing the seamy underside of the investing game. _ books and articles exposed penny-stock scammers and greedy chief executives. _ 2002 book,
"martha inc. : the incredible story of martha stewart living omnimedia," was made into a television movie starring cybill shepherd. about 16 years earlier, _byron had written about the fumbling early attempts by executives at time inc. to adapt to a rapidly shifting media landscape. _ 1986
book, "the fanciest dive: what happened when the giant media empire of time/life leaped without looking into the age of high-tech," foreshadowed the equally disastrous merger of time warner and aol a decade and a half later. indeed, the time inc. tale has held up well. in a 2008 column,
joe nocera of the new york times ranked it among the best nonfiction business books of recent decades. _byron's 1992 book, "skin tight: the bizarre story of guess v. jordache," looked at the fierce rivalry of two blue-jean powerhouses. "_ was dogged in _ journalism," said joni evans, an
editor and literary agent who had represented _byron. "chris was passionate about _ subjects and never let go. "christopher michael byron was born on dec. 27, 1944, in washington, d. c. _ parents, edward armour byron and the former ella katherine mccune, both worked in radio and
later in television — _ father as a producer, _ mother as an actress — giving _byron an early taste of life in the media. after dropping out of stamford high school in connecticut in 1962, _byron served in the navy for two years before talking _ way into yale, even though _ did not have a
high school diploma. _ graduated with honors in 1968. that same year, _ married maria los, whom _ had met while _ was at yale and _ was a student at connecticut college in new london. they divorced last year. besides _ daughter katy, a managing news editor at snapchat, _byron is
survived by another daughter, jana byron, a maritime lawyer; a son, nicholas byron, an artist; and a brother, kevin byron, a nature photographer. after earning _ first bylines at the hour in norwalk, conn. , and graduating from columbia law school in 1972, _byron joined the staff of time
magazine. _ was later a foreign correspondent for time in bonn, germany, and london. after stints at forbes, new york magazine and esquire, _byron wrote a financial column for the new york observer from 1995 to 2001.